BIOINFORMATICS APPLICATIONS NOTE Vol. 25 no. 5 2009, pages 692–694
doi:10.1093/bioinformatics/btp042

Databases and ontologies

DSSTox chemical-index files for exposure-related experiments in ArrayExpress and Gene Expression Omnibus: enabling toxico-chemogenomics data linkages

ClarLynda R. Williams-DeVane[1,*], Maritja A. Wolf[2] and Ann M. Richard[1]

[1]National Center for Computational Toxicology, Office of Research and Development, US EPA and [2]Lockheed Martin, Research Triangle Park, NC 27711, USA

Received on October 31, 2008; revised on January 12, 2009; accepted on January 18, 2009

Advance Access publication January 21, 2009

Associate Editor: Alex Bateman

ABSTRACT

Summary: The Distributed Structure-Searchable Toxicity (DSSTox) ARYEXP and GEOGSE files are newly published, structure-annotated files of the chemical-associated and chemical exposure-related summary experimental content contained in the ArrayExpress Repository and Gene Expression Omnibus (GEO) Series (based on data extracted on September 20, 2008). ARYEXP and GEOGSE contain 887 and 1064 unique chemical substances mapped to 1835 and 2381 chemical exposure-related experiment accession IDs, respectively. The standardized files allow one to assess, compare and search the chemical content in each resource, in the context of the larger DSSTox toxicology data network, as well as across large public cheminformatics resources such as PubChem (http://pubchem.ncbi.nlm.nih.gov).

Availability: Data files and documentation may be accessed online at http://epa.gov/ncct/dsstox/.

Contact: williams.clarlynda@epa.gov

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

In recent years, the number of publicly available gene expression, toxicology and cheminformatics resources with the potential to support toxicogenomics investigation has grown considerably (http://www.microarryaworld.com/DatabasePage.html; Richard *et al*., 2008; C.R.Williams-DeVane *et al*., manuscript submitted). These trends are encouraging aggregation and use of data in a much broader context, spanning domains of inquiry in relation to toxicology, chemistry and genomics (Waters *et al*., 2008).

The European Bioinformatics Institute's (EBI) ArrayExpress Repository (http://www.ebi.ac.uk/microarray-as/ae/) and the National Center for Biotechnology Information's (NCBI) Gene Expression Omnibus (GEO) (http://www.ncbi.nlm.nih.gov/geo/) are the two main public repositories for gene expression experiments associated with the published scientific literature. Although they each support MIAME-compliant submissions (i.e. adhering to guidelines for Minimum Information about Microarray Experiments; http://www.mged.org/Workgroups/MIAME/miame.html), neither resource has standardized requirements for reporting of chemical information associated with submitter-deposited microarray experiments. As a result, not only has it been difficult to assess the chemical-related content within these resources, but also microarray data have been effectively isolated from rapidly growing public sources of chemically indexed information pertaining to toxicology (Richard *et al*., 2006).

We report here the publication of chemical-index files for experimental content in the ArrayExpress Repository and GEO Series (data extracted on September 20, 2008), in association with the Environmental Protection Agency's (EPA) Distributed Structure-Searchable Toxicity (DSSTox) Data Network project (http://www.epa.gov/ncct/dsstox/) (Supplementary References).

2 DATABASE METHODS AND COMPONENTS

2.1 DSSTox

The DSSTox project publishes high-quality, standardized chemical structure toxicity data files pertaining to high-interest chemicals for environmental toxicology and of potential use for structure–activity relationship (SAR) modeling. The DSSTox website offers documentation, freely downloadable structure data files (SDF) and tabular data files (.xls) for each published Data File (http://www.epa.gov/ncct/dsstox/DataFiles.html). A unique aspect of this effort is the quality annotation, review and representation of chemical information both in terms of a unique mapping to a curated chemical structure, as well as at the generic test substance level (similar to Chemical Abstracts Service (CAS) Registry Number distinctions)—see http://www.epa.gov/ncct/dsstox/MoreonStandardChemFields.html. The current DSSTox inventory contains over 8000 unique chemicals and has been incorporated into the online DSSTox Structure-Browser (http://www.epa.gov/dsstox_structurebrowser/), the NCBI PubChem (http://pubchem.ncbi.nlm.nih.gov/) inventory containing millions of searchable chemical structures and thousands of bioassays, ChemSpider (http://chemspider.com/) containing millions more chemical structures, properties and linkages and the new EPA Aggregated Computational Toxicology Resource (ACToR) database (http://www.epa.gov/actor/) providing searchability and comparative read-across for over 200 chemical inventories specifically pertaining to environmental toxicology.

*To whom correspondence should be addressed.

2.2 ArrayExpress

Since the public launch of ArrayExpress in 2002, the repository has grown to more than 6500 experiments. Public microarray data in ArrayExpress are available for browsing and querying across a wide range of experiment properties, including array type, submitter, species, MIAME score, etc. and complete datasets or subsets can be retrieved (http://www.ebi.ac.uk/microarray-as/aer/entry). Although the TOXM label is available to designate toxicogenomics experiments, it is rarely used and represents a very small portion of the total ArrayExpress chemical–experiment inventory. Chemical involvement is primarily indicated by non-standard, error-prone chemical names or abbreviations included in free-text user description fields, and these are very rarely accompanied by chemical identifiers such as CAS or Chemical Entities of Biological Interest (ChEBI) numbers. Where ChEBI identifiers are used (∼20 instances), these have been recently cross-referenced within the ChEBI system (http://www.ebi.ac.uk/chebi/).

2.3 GEO series

A GSE is a GEO Series Accession ID (e.g. GSE5594) that defines a set of related samples considered to be part of a single experiment and, for present purposes, most closely (or precisely) corresponds to an ArrayExpress Accession ID (e.g. E-GEOD-5594). GEO currently contains over 9900 Series Accession entries and more than 2000 curated GEO Datasets. However, fewer than 6500 of the Series Accession entries could be programmatically extracted due to a backlog in the GEO curation process at the time of annotation. Chemical information is most often located in GEO Series records as chemical names or abbreviations in the Summary field (a user-submitted, free-text description field), but in some cases this information only was provided in the 'Title' or 'Samples' field. A handful of GEO records contained chemical identifiers such as CAS, but the quality of chemical annotation across GEO, in general, was poor and in some cases absent entirely.

2.4 Methods: ARYEXP_Aux and GEOGSE_Aux

Recent additions to ArrayExpress include new portals for programmatic access where users can query and download data in a systematic manner from the ArrayExpress FTP site (http://www.ebi.ac.uk/microarray/doc/help/programmatic_access.html).
To create the ARYEXP auxiliary file (ARYEXP_Aux), content was extracted using Perl scripts from the programmatic access FTP site in XML format. Similarly, the creation of GEOGSE_Aux involved first programmatically accessing GEO to retrieve all current experimental descriptions associated with Series records (http://www.ncbi.nlm.nih.gov/projects/geo/info/geo_paccess.html). Entrez tools were used to generate an XML document containing a summary of each of the GEO Series experiments currently curated in the GEO Data Sets Accession ID (GDS) system. A series of Perl scripts were developed to parse both the ArrayExpress and GEO XML documents and to retrieve records with probable chemical association. The resulting experiment descriptions were evaluated and verified manually, and chemical information was extracted and subsequently underwent stringent review and annotation according to DSSTox procedures (http://www.epa.gov/ncct/dsstox/ChemicalInfQAProcedures.html).

Each of the resulting DSSTox files, ARYEXP_Aux and GEOGSE_Aux, is a chemical–experiment pair index (one record per chemical per experiment). Each file also contains the full complement of 20 DSSTox Standard Chemical Fields and an additional 14 Standard Genomics Fields (including URL field linking to experiment accession IDs) to allow cross comparisons of ArrayExpress and GEO content similarly indexed by DSSTox (Supplementary Tables 1 and 2). In addition, ARYEXP_Aux contains an additional set of 30 experimental description fields specific to ArrayExpress (Supplementary Table 3), and GEOGSE_Aux contains an additional set of four experimental description fields specific to GEO Series (Supplementary Table 4). Further details of chemical-indexing and data extraction methods, comparison of ArrayExpress and GEO chemical–experimental summary content and assessment of toxicologically relevant chemical content are provided elsewhere (C.R.Williams-Devane, manuscript submitted; Supplementary References).

2.5 Methods: ARYEXP and GEOGSE

Of greatest toxicogenomics and SAR interest are those experiments for which a chemical treatment and the resulting gene expression changes are the primary focus of the experiment. We introduced the Standard Genomics Field, Chemical_StudyType, to annotate the purpose of the chemical in the experiment, which could include uses such as 'Treatment', 'Vehicle', 'Reference', 'Media', etc. The main DSSTox files, ARYEXP and GEOGSE, pertain only to the 'Treatment' category of ArrayExpress and GEO Series experiments and contain one record per unique chemical substance. In ARYEXP and GEOGSE, one chemical substance can map to one or more experiments in GEO or ArrayExpress. Hence, unlike the Auxilliary files, where records map to individual experiments, the main DSSTox structure-index files do not contain summary details of particular experiments. Rather, these files contain DSSTox Standard Chemical Fields, Chemical_StudyType (which can be Treatment AND other conditions), one or more Experiment Accession IDs and the corresponding URLs to ArrayExpress or GEO Series Experiment Summary pages.

3 CONCLUSIONS AND PERSPECTIVES

The ARYEXP_Aux file contains a total of 2365 chemical–experiment records (with 44 total source fields), corresponding to 1011 unique chemical substances. Of these 2365 chemical–experiment pairs, 1835 were identified as 'Treatment' and these map to 887 unique chemical records in the ARYEXP file. Similarly, the GEOGSE_Aux file contains a total of 2381 chemical–experiment records (with 18 total source fields), corresponding to 1064 unique chemical substances. Of these 2381 chemical–experiment pairs, 2134 were identified as 'Treatment' and these map to 1014 unique chemical records in the GEOGSI file. These numbers indicate that the exposure-related experimental content in these two public resources covers a significant range of chemicals of potential interest and utility for toxicogenomics investigations. All four files are available for download from the DSSTox website (http://www.epa.gov/ncct/dsstox/).

The ARYEXP and GEOGSE chemical-index files, with associated ArrayExpress Experiment Accession ID URLs, have been incorporated into the DSSTox Structure-Browser, ACToR and PubChem. This enables ArrayExpress and GEO Series experiments to be structure located from these public resources, creating the

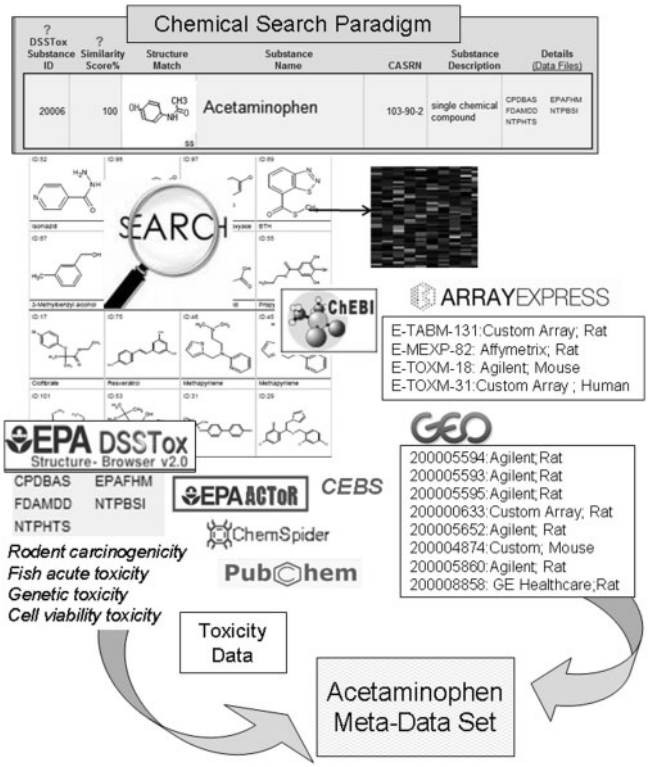

Fig. 1. Illustration showing linkages from DSSTox Structure-Browser to experiments from ARYEXP and GEOGSE, linked to various chemically indexed resources of bioassay and toxicity data for constructing a meta-dataset for acetaminophen; links to actual microarray data are provided by GEO or ArrayExpress (dashed arrows indicate future linkages).

capability to query multiple domains of data by chemical structure. Figure 1 illustrates this capability with a chemical structure search of acetaminophen. Summary toxicology and microarray results are provided or located through the DSSTox structure search, and other domain results are provided through linkages with PubChem, ACToR and ChemSpider. With these new capabilities, a meta-dataset on a particular chemical or family of structurally similar chemicals could be constructed for further analysis.

The DSSTox ARYEXP and GEOGSE chemical-index files have been deposited within PubChem such that chemicals can be located by keyword search under 'PubChem Substance' on the main search page (e.g. ARYEXP, ArrayExpress, etc.). From the PubChem Substance results page, a user can link directly to the chemical-associated experimental accession ID summary pages in GEO and ArrayExpress. Likewise, through chemical linkages, these microarray data could be placed in a much larger data and chemical context (including linkage to data for structurally and biologically similar chemicals) within PubChem. ARYEXP and GEOGSE auxiliary data files contain summary experimental factors (Supplementary Tables 3 and 4), but do not contain actual microarray data or annotation files. To encompass these data types, integration of ArrayExpress and GEO into the Chemical Effects in Biological Systems (CEBS; http://cebs.niehs.nih.gov/) toxicogenomics database is (Waters *et al*., 2008). An automated process of porting microarray data and annotation files directly from ArrayExpress and GEO to CEBS is to be implemented with chemical annotation handled in collaboration with the DSSTox project. Recommendations for chemical standards for microarray experiments are being forwarded to the MIBBI (Minimum Information for Biological and Biomedical Investigations) project (http://www.mibbi.org/).

Better coordination between NCBI's GEO and PubChem projects is needed. Likewise, EBI's ChEBI and ArrayExpress projects have recently improved their coordination. The present effort has chemically annotated a large portion of the current inventories of ArrayExpress Repository and GEO Series, although more efficient mechanisms for updates must be instituted. The preferred solution is to incorporate standard chemical reporting requirements into these resources directly. The present effort charts a path forward and will be used to encourage implementation of these changes. ArrayExpress currently has the ability to adequately capture most of the required chemical information; however, depositors are not doing so. We recommend that GEO and ArrayExpress, in coordination with projects such as MIBBI and DSSTox, each adopt formal requirements for a minimum level of chemical annotation in relation to experiments (e.g. valid chemical name, Chemical_StudyType). Once repositories and depositors recognize the importance and enhanced capabilities to be gained from chemical annotation and indexing, and make it a priority, the possibility to more fully integrate and utilize existing data in toxicogenomics studies can be realized. *Updates to DSSTox GEOGSE and ARYEXP files by current means are scheduled for February 2009.*

ACKNOWLEDGEMENTS

This manuscript was approved by the US EPA's National Center for Computational Toxicology for publication; the contents do not necessarily reflect the views and policies of the EPA and mention of trade names or commercial products does not constitute endorsement or recommendation for use.

Funding: NCSU/EPA Cooperative Training Program in Environmental Sciences Research, Training Agreement CT833235-01-0 with North Carolina State University (to C.R.W.).

Conflict of Interest: none declared.

REFERENCES

Richard,A. *et al*. (2008) Toxicity data informatics: supporting a new paradigm for toxicity prediction. *Tox. Mech. Meth.*, **18**, 103–118.

Richard,A.M. *et al*. (2006) Chemical structure indexing of toxicity data on the Internet. *Curr. Opin. Drug Discov. Dev.*, **9**, 314–325.

Waters,M. *et al*. (2008) CEBS—chemical effects in biological systems: a public data repository integrating study design and toxicity data with microarray and proteomics data. *Nucleic Acids Res.*, **36**, D892–D900.